CTAB-GAN: Effective Table Data Synthesizing

Year: 2021 | Citations: 248 | Authors: Zilong Zhao, A. Kunar, H. V. D. Scheer, R. Birke, L. Chen

Abstract

While data sharing is crucial for knowledge development, privacy concerns and strict regulation (e.g., European General Data Protection Regulation (GDPR)) unfortunately limit its full effectiveness. Synthetic tabular data emerges as an alternative to enable data sharing while fulfilling regulatory and privacy constraints. The state-of-the-art tabular data synthesizers draw methodologies from generative Adversarial Networks (GAN) and address two main data types in the industry, i.e., continuous and categorical. In this paper, we develop CTAB-GAN, a novel conditional table GAN architecture that can effectively model diverse data types, including a mix of continuous and categorical variables. Moreover, we address data imbalance and long-tail issues, i.e., certain variables have drastic frequency differences across large values. To achieve those aims, we first introduce the information loss and classification loss to the conditional GAN. Secondly, we design a novel conditional vector, which efficiently encodes the mixed data type and skewed distribution of data variable. We extensively evaluate CTAB-GAN with the state of the art GANs that generate synthetic tables, in terms of data similarity and analysis utility. The results on five datasets show that the synthetic data of CTAB-GAN remarkably resembles the real data for all three types of variables and results into higher accuracy for five machine learning algorithms, by up to 17%.